An Overview of Search Engine

Hai-Yang Xu
Dev Lead of Search Technology Center
Microsoft Research Asia
haixu@microsoft.com
July 24, 2007
Outline

- History of Search Engine
- Difference Between Software and Service
- Architecture of Search Engine
- 5 Tips On Optimizing Search Engine
- 3 Secrets On Implementing Search Engine
History of Search Engine

- Personal or Academic Site (1993-1996)
- Internet Portal (1996-1999)
- Technology Provider (1999-2002)
- Search Portal (2002- )
Personal or Academic Site (1993-1996)

- Archie
- WebCrawler
- Lycos
- Excite
- Yahoo
Internet Portal (1996-1999)

- Yahoo!
- Lycos
- Excite
- Infoseek
Technology Provider (1999-2002)

- AltaVista
- Inkotomi
- Fast/AlltheWeb
- Google
- Goto/Overture
Search Portal (2002-)

- Google
- Yahoo
- MSN
- ASK
Lessons From The Past

- Technology is the biggest challenge
- Search engine always is an important application of Internet
- Search engine can always be developed better
Architecture of Search Engine

- **URL DB**
- **Crawler**
- **Page DB**
- **Index**

**Flow:**
- URL DB → Crawler → Page DB → Index

**Query Flow:**
- Query → Search → Inverted Index

**Result Page Flow:**
- Search → Result Page
Difference Between Software and Service

- Product vs. Experience
- Feature vs. Refinement
- Develop vs. Operate
- Release vs. Serve
- Code vs. Parameter
- Update vs. Tune
- Bug Free vs. Optimal
Crawler

- Crawling is more difficult than what you think
- Stability in downloading, computation and storage
- Scalability
- High Performance
Index

- Performance
- Content Analysis
  - tf*idf
  - html tag
  - html visual information
- Link Analysis
  - PageRank,
  - Spam
Search

- Huge Traffic
- Huge Index Data
- Complicated Computation
- Very Large Server Cluster
Engineering Problem of Search Engine

- Intellectual Problem
  - Optimization
- Non-intellectual Problem
  - Implementation
5 Tips On Optimizing Search Engine

- Define problem from user perspective
- System-level thinking
- Feature is more important than classification method
- Tradeoff
- Combine several simple solutions to a powerful solution
Non-intellectual Problem

- Architecture
- High Performance
3 Secrets On Implementing Search Engine

- Cache
- Signature
- Hash Table
Cache

- What is cache
- What to do with cache is more important than how to cache
- Search result page cache
Cache (Cont.)

- Front-end Cache & Back-end Cache
- Caching Merged Index & Caching Raw Index
- Caching Index & Caching Display Information
- Caching everything
Cache (cont.)

- Cache them before search
- Cache In Disk
- 2 Terms Cache
Signature

- What is signature
- What problems can benefit from signature trick
- Signature algorithm
- Probability of conflict
Hash Table

- What is hash table
- Implementation of hash table using signature
Inverted Index

- Document is an Index by doc
- Inverted Index is an Index by term
- Inverted Index just is a hash table of terms
Query Statistics

Problem:

- From query log file, we want to get frequency of each query

Solution 1:

- Sort query log file, then count each query

Solution 2:

- Hash table
O(n) Sort

Problem:
- Sort student record according to examination score

Solution 1:
- qsort(), O(n*\log n)

Solution 2:
- Hash table, O(n)
De-duplicate URLs

Problem:
- De-duplicate URLs with same content

Solution 1:
- Sort, then compare

Solution 2:
- Hash table
Set Operations

- A*B
- A+B
- A-B
- B-A
Page Storage Architecture

1. Crawl → Page DB → Index

2. Crawl → Page Distributor → Index

Page DB → Page DB → Page DB
Index Architecture
Search Architecture
Crawling Architecture

How about this solution:

WRONG!
Crawling Architecture (Cont.)
Summary

- History of Search Engine
- Difference Between Software and Service
- Architecture of Search Engine
- 5 Tips On Optimizing Search Engine
- 3 Secrets On Implementing Search Engine
Thank you!
Q&A